Historic, Archive Document

Do not assume content reflects current scientific knowledge, policies, or practices.

Reserve
1.942
C32St2
1934

UNITED STATES
DEPARTMENT OF AGRICULTURE
LIBRARY

Reserve
BOOK NUMBER 833181

1.942
C32St2
1934

UNITED STATES DEPARTMENT OF AGRICULTURE
Bureau of Agricultural Economics

October, 1934

Analysis of Variability in Staple-Length Observations on Press-Box and Cut Samples of Cotton

- - - - - - -

By

F. H. Harper

LIBRARY RECEIVED MAR 23 1953 U. S. DEPT. OF AGRICULTURE

- - - - - - -

Increasingly keen interest has been manifested in comparatively recent years in statistical interpretations of numerical data. This interest has been responsible for the development of numerous analytical methods and for their application to the study of variability in data pertaining to many fields.

Present-day statisticians, as well as practical and theoretical economists, have borrowed extensively from contributions made by biologists and sociologists, and, among other uses, they have appropriately adapted certain probability and error theories in their measurements of relationships and in their analysis of variation. These theories have been applied to some extent also in the appraisal of the magnitude of variances, and they have been found useful in deciding on the probabilities of whether or not different series of observations and their means represent unstratified populations.

Economists, agronomists, and other workers have found it impossible to solve many of their problems to their own satisfaction by the mere application of correlation methods and by the interpretation

of variation from results obtained through application of the more commonly-used variability and dispersion measures. Need is sometimes felt for measures that indicate the relative magnitudes of contributions made from different detected sources to total variability. This is particularly true in those analyses of differences in paired and replicate series of observations which necessitate, in addition to a determination of the magnitudes of different parts of variability contributing to the total, an appraisal of differences both between and within the series. It is for this reason, at least in part, that present-day statistical analysts have found it convenient to employ more comprehensive techniques in the interpretation of variation in their sample data, and in the evaluation of inherent, or characteristic, differences in paired and replicate series of observations.

It seems logical in analyzing the total variability in paired and replicate series of data to assume that the different sources of variability must first be detected before any reliable estimate can be made as to the relative importance or significance of the contributing parts. Unless this is done, it is difficult, indeed, if not impossible, to logically appraise the relative importance of the different parts of variability. Fortunately, the present day analyst is now able to proceed in this way and thus ~~to more comprehensively~~ interpret his data. He can do so either (1) by separating into its component parts the summation of squares of the differences between individual observations and their common mean and then proceeding to calculate the different parts of average squared variability, or (2) by calculating the squares of the

standard deviations by one of the ordinary methods and then isolating the individual parts of variability. Choice of method depends somewhat upon the characteristics of the data themselves, the sources from which variability is contributed, and the use that is to be made of the derived results.

It has been quite appropriately suggested by Dr. R. A. Fisher, eminent English statistician, working with agronomic data at the Rothamsted Experimental Station, Harpenden, England, that the term "variance" 1/ be applied to the standard deviation squared, and this usage of the term has been quite generally adopted. Investigators who have applied the analysis of variance method to the interpretation of variability have been especially impressed with the possibility of its more extensive adaptation and usefulness and the convenience with which it may be used in many instances to analyze squared variability in measuring differences between averages and within series of paired and replicate observations.

In this report an analysis is made of differences in the classification of press-box and cut samples of cotton taken from the same bales. 2/ The basic data represent actual staple-length designations by government classers and are a part of a larger mass of data compiled and analyzed, ~~all of which is available for inspection, together with the results of the analysis.~~ The two sets of samples

---

1/ The credit for introduction of the statistical technique known as "Analysis of Variance" belongs to Dr. R. A. Fisher and his associates. It was published in 1923 (Fisher, R. A., and Mackenzie, W. A., Jour. Agri. Sci., Vol. 13, Part 3, July, 1923, pp. 311-320), and since that time, having been used in statistical studies of numerous descriptions, has been found to be the most appropriate method yet suggested for many mathematical inquiries into the probability of significance of differences.

2/ Press-box samples are those taken from the gin press-box while the bales ~~were~~ in process of being ginned, whereas cut samples are those taken from the same bales after they had been pressed and tied.

were classed by the same classers and under as nearly uniform conditions as practicable to maintain at the time. 3/

The purpose of the analysis herein described is not primarily to direct attention to the basic data, but rather to illustrate the application of a method of statistical procedure that has been successfully used in the interpretation of differences in classification of press-box and cut samples of cotton, by which other series of observations falling in the same general category may be analyzed and the differences between them and within them appropriately evaluated. An additional statistical procedure is described in this paper by which relative magnitudes of the different parts of squared variability are logically appraised. 3/ If it were not for the desirability of separately evaluating the different component parts of squared variability, the derivation of the *t* value would serve the same purpose in the interpretation of differences

---

2/ (cont'd.)
During the 1930-31 season, the Grade and Staple Statistics Section of the Division of Cotton Marketing, U. S. Department of Agriculture, conducted a study for the purpose of obtaining information on the possible effects of gin compression on staple-length designation of cotton samples. A similar study was conducted on a smaller scale during the 1929-30 season. Several office reports have been prepared in which are presented in detail different phases of the results of these studies. The classification data used in this report are a part of those procured in connection with the study made in 1930-31.
The analysis of variance method and the procedure which permits variability to be separated into component parts free from estimates of error was introduced into the Division of Cotton Marketing in October, 1930. Since that time they have been found increasingly useful in the interpretation of cotton classing and other variability.

3/ Kemp, W. B., Jour. Amer. Stat. Assn., Vol. XXIX, No. 186, June, 1934, p. 147.

between means as is served by the z value in the analysis of variance.

It is to be realized at the outset that lack of agreement in magnitude of paired observations on staple length of cotton samples may result in "bias" as well as the inevitable "spread" in the distributions attributable to the lack of agreement. For convenience, therefore, the former term is used to indicate that part of the total variability contributed by the net difference, or difference on the whole, between different series of observations. It is obvious, then, that the bias is necessarily represented in the "spread" and is a part of it. Usage of the term "bias" is not to be understood, however, as implying that observations in one series differ consistently from those in another series.

The bias may quite generally account for only a part of the variability caused by lack of agreement in magnitude of paired observations, so that one of the principal problems in analyzing differences in classing is to make an appraisal of the bias in order that there may also be available a measure of the variability attributable to "compensating tendencies." To this latter measure there is herein applied the term "error," which, in some respects a residual, is indicative of the variability within series of paired observations that is contributed, in addition to bias, by the failure of these observations to agree in magnitude.

The necessity for separately evaluating bias and error as interpreted in this discussion is fully realized when an attempt is made to compare the measures of variability contributed from the two

sources. It would not be logical, of course, to express the bias and error in terms different from that representing "spread" and then attempt comparisons. Furthermore, the "spread" includes both bias and error, and the two must be treated as distinct and separate measures in order to show the magnitude of each and to indicate their relative importance. The analysis of variance method is used in determining probability of significance of differences between measures of average squared variability, and the method suggested by Dr. Kemp is used in showing the relative magnitudes of the component parts of total squared variability. Technique involved in the application of these methods is illustrated by the analysis herein made.

The following equations, 4/ representing technique that has been applied in studying the differences in classifications of press-box and cut samples of cotton taken from the same bales, will indicate the nature of the calculations and the fundamental mathematical principles underlying them. 5/

1. Correction factor (c) = $\sum(x + y)\ \frac{\sum(x + y)}{2n}$, or $\frac{\left[\sum(x + y)\right]^2}{2n}$, in which "x" and "y" represent the individual observations in the two series, and "n" the number of observations in ~~each of~~ the series, which is the same, of course, as the number of paired observations when each is present in duplicate.

2. Total summation of squares, or total squared variability,

$$= \sum x^2 + \sum y^2 - c$$

4/ The analytical technique represented by these equations has been applied also to studies of differences in classifications of identical samples by two or more classers and to studies of differences in classifications of different samples from the same bales. The method is applicable in some instances to the interpretation of variation in prices and other economic data.

5/ Another method (unpublished) of analysis, furnishing comprehensive interpretations of variation and mean differences, has been suggested by O. T. Weaver and successfully used by him.

3. Bias, or squared variability attributable to the difference between the summations and, consequently, the means of x and y,

$$= \frac{(\Sigma x)^2}{n} + \frac{(\Sigma y)^2}{n} - c$$

4. "Sample" variability, or the squared variability attributable to differences in magnitudes of successive x + y summations, or to accepted differences within the series upon which the observations are made, which would necessarily result in differences in magnitudes of successive x + y summations,

$$= \frac{\Sigma(x + y)^2}{2} - c$$

5. Error = total squared variability (equation 2) - sample variability (equation 4) - bias (equation 3)

The data in table 1, in which there is an average difference of 0.7 of one-sixteenth of an inch between the two series of paired observations, are compared and the differences between and within the series interpreted. Application of the equations is made in the interpretations, and the calculations suggested by them are illustrated.

Table 1.- Distribution of staple-length designations of 10 paired press-box and cut samples of cotton 1/

Sample number	Staple-length designation (sixteenths of an inch)		x + y
	x (press-box samples)	y (cut samples)	
1	13	12	25
2	13	15	28
3	14	15	29
4	14	14	28
5	15	17	32
6	16	17	33
7	16	16	32
8	17	16	33
9	18	20	38
10	19	20	39
Total	155	162	317
Mean 2/	15.5	16.2	15.85 3/

1/ Selected from unpublished data. Press-box samples are those taken from the gin press-box while the bales were in process of being ginned, whereas cut samples are those taken from the same bales after they had been pressed and tied.

(continued on next page)

2/ Calculated for comparative purposes on the lower limits of the staple-length groups. Means based on midpoints may be obtained by adding 0.5.

3/ Representing the common mean of the two series of observations.

1. Correction factor = 317 x 15.85 = 5024.45

2. Total squared variability = 2441 + 2680 - 5024.45 = 96.55

3. Bias, or x - y squared variability, = $\frac{(155)^2}{10} + \frac{(162)^2}{10} -$ 5024.45 = 2.45

4. "Sample" variability = $\frac{10225}{2}$ - 5024.45 = 88.05

5. Error = 96.55 - 88.05 - 2.45 = 6.05

Correction factor. - This value is determined, as the equation indicates, by obtaining the product of the summation of observations and their mean, which is the equivalent of the quotient resulting from dividing the square of the summation of observations (square of 317) by the number of observations (20). It represents the difference between the summation of the squared observations and the summation of the squares of deviations from the common mean, 15.85, and it may be expressed as follows:

$$\text{Correction factor} = \Sigma x^2 + \Sigma y^2 - \left[\Sigma(x - \overline{x + y})^2 + \Sigma(y - \overline{x + y})^2\right],$$

in which $\overline{x + y}$ represents the common mean of the x and y observations.

Total squared variability. - This summation of squares, as calculated, is the difference between the summation of the squares of all observations and the correction factor, the correction factor being subtracted because it constitutes the difference between the summation of squares of individual observations and the summation of squares of deviations from the common mean. It is the equivalent of the total of the squares of deviations of all observations in the two series from the common mean, as the following equation indicates: Total squared variability = $\Sigma(x - \overline{x + y})^2 + \Sigma(y - \overline{x + y})^2$, in which x and y are

individual observations and $\overline{x + y}$ the common mean. The measure of total squared variability may be obtained by squaring the deviations of individual observations in table 1 from their common mean, 15.85, and then summating the squares.

When paired observations are of the same magnitude, the total of their squares is one-half as great as the square of their summation and equal to the product of their mean and summation. When paired observations are not of the same magnitude, the total of their squares exceeds one-half the square of their summation by an amount equal to one-half the square of the difference between the two observations; ~~and~~ it exceeds the product of their mean and summation, also, to the extent of one-half the square of this difference.

Bias. - As will be observed by the equation already presented, and by the calculations following table 1, this measure of variability, containing, as derived, one part of error (0.6722, column 4 of table 2), since there is one degree of freedom (column 2 of table 2), is contributed by differences, on the whole, between the series. In the calculations presented, it is the quantity, plus indicated error, that remains after the correction factor has been subtracted from the summation of the quotients obtained by dividing the squares of the summations of both the x and y series by the number of observations in each series.

When calculated by the method described, it is the equivalent of the product of the total number of observations and the square of one-half the difference between the mean of the x series of observations

and the mean of the y series, one-half of the difference between these two means being the same as the total difference between the common mean and the mean of each series. This may be expressed by the following equation:

Bias $= 2n\left(\frac{\bar{x} - \bar{y}}{2}\right)^2$, in which "n", as already stated, represents the number of paired observations.

The calculated means of the two series of observations shown in table 1 are 15.5 and 16.2, respectively, the difference between them being 0.7, the equivalent of the arithmetic mean of the algebraic summation of the differences between the paired observations. It is obvious that the algebraic summation of differences between individual observations in both series and their common mean, 15.85, is necessarily 0.

Since there is an average difference of 0.7 between the two series of paired observations, the quantity 2.45 can be obtained also by squaring this average difference, multiplying by the number of paired observations, and then dividing by 2. It will be realized, of course, that the product of the square of this average difference and the number of paired observations is twice as great as the product of the square of one-half the average difference and the total number of x and y observations, and that it is four times as great as the product of the square of one-half the average difference and the number of paired observations. Comprehension of these relationships between magnitudes of products makes it possible to comprehend more readily the measure of bias.

<u>Sample variability</u>. - This measure, containing parts of error (0.6722) as indicated by the degrees of freedom in table 2, is calcu-

lated in the analysis of data in table 1 by squaring the summations of pairs of observations, summating the squares, dividing by 2, and then subtracting the calculated correction factor. Sample variability as herein calculated is contributed by differences in the magnitudes of successive $x + y$ summations (See table 1). These differences in $x + y$ summations are present when successive observations in the individual series are of unlike magnitude and paired observations are in perfect agreement; and they may be present either when successive observations are of unlike magnitude and paired observations are not in perfect agreement, or when successive observations in one series are of like magnitude and corresponding observations in the paired series are not in agreement.

The measure may be calculated also by squaring the deviations of observations in each individual series from their mean, summating the squares, and subtracting the error, which, as will be shown by the following explanation, may be calculated independent of the correction factor, total squared variability, and "sample" variability.

<u>Error</u>. - For purposes of the analysis of variance, a measure of error may be obtained by subtraction, it being the quantity remaining, as indicated by the equations and by the calculations following table 1, after measures of variability calculated for bias and "sample" have been subtracted from total squared variability. Error, ~~indicative of compensating "spread",~~ is contributed by lack of agreement in the magnitude of observations, and it is in addition to bias, or net

difference, which is also contributed by lack of agreement. 6/ It will be understood, of course, that there is no bias unless the differences in series of observations are non-compensating, so that the summation and average of one series are larger or smaller than the summation and average of another.

Problems often confronting the analyst when observations are not in agreement in respect to magnitude and when averages of series differ are, as already indicated, the determination of the probability of significance of differences between variances, and the separation of total squared variability into its component parts. These are essential if differences are to be properly evaluated and if the most desirable comparisons are to be made between the different parts of squared variability.

In applying the analysis of variance method, the desired measure for error may be determined independent of the correction factor, total squared variability, and "sample" variability. (A convenient advantage is thus afforded in checking computations.) This calculation is made by determining the difference between magnitudes of each pair of matched observations, squaring these differences, dividing by 2, summating, and then subtracting the bias. Instead of dividing the square of each difference by 2 and then summating, the squares may be summated and the total divided by 2.

With the total squared variability separated into designated component parts, each containing as many parts of average squared

---

6/ It has been observed in the classification of press-box and cut samples representing the same bales, and also in the classification of identical samples by the same and different classers, that there is frequently inconsistent variation in the distribution of staple-length observations. Tolerance in classing may possibly account for an appreciable part of this inconsistency, and the differences themselves between magnitudes of paired staple-length observations may be due in part to actual differences in the cotton upon which observations are made.

variability for error as are indicated by the corresponding degrees of freedom, a further calculation consists of determining whether one estimate of squared variability obtained from n_1 degrees of freedom 7/ differs significantly from another estimate of squared variability obtained from n_2 degrees of freedom. Fortunately, the problem is simple, it being only necessary to calculate the z value 8/ equal to half the difference between the natural logarithms of the two derived measures of average squared variability, or to the difference between the natural logarithms of the corresponding standard deviations (i.e., the difference between the natural logarithms of the square root of the measures of average squared variability).

The values in column 4 of table 2 represent the calculated averages of squared variability, obtained by dividing the squared measures in column 3 by the degrees of freedom in column 2. Then, as Fisher has explained, if P represents the probability of exceeding the calculated z value by mere chance, it becomes possible to obtain the value of z corresponding to different values of P, n_1, and n_2. 9/

7/ The term "degrees of freedom" is used in stating the number of series of observations or the number of observations within series that may be free to vary from any single series or observation. In Fisher's table of 5 percent points, n_1 corresponds to the larger variance as calculated in tables 2 and 4.

8/ The distribution of this z value is closely related in principle to the distribution of z values worked out by "Student" and Pearson. Fisher's z value is equal to one-half of the natural logarithm of the quotient obtained by dividing one average squared variability, such as presented in column 4 of table 2, by another. It is calculated in this report by determining the difference between one-half the natural logarithms of two derived estimates of variance, which is the equivalent of one-half the difference between two such logarithms. A comprehensive explanation of analytical procedure is presented by C. H. Goulden in a paper entitled "Application of the Variance Analysis to Experiments in Cereal Chemistry" (Cereal Chemistry, Vol. IX, No. 3, May, 1932, pp. 239-260).

9/ Readers may be interested in an article by T. Eden, Tea Research Institute of Ceylon, and F. Yates, Rothamsted Experimental Station, entitled "On the Validity of Fisher's Z Test When Applied to an Actual Example of Non-normal Data " (Jour. Agr. Sci., Vol. 23, Part 1, January, 1933, pp.6-17.)

The following table, based on table 1 and the results of the analysis of data contained therein, illustrates the procedure by which it may be determined whether or not the measure of average squared variability obtained from n_1 degrees of freedom is significantly greater than that obtained from n_2 degrees of freedom.

Table 2. - Sources of variability, degrees of freedom, and measures of variability contributed from specified detected sources

1	2	3	4	5
Source of squared variability	Degrees of freedom	Squared variability (summation of squares)	Average squared variability 1/	$\frac{1}{2}$ log.$_e$ 2/
Bias	1	2.45	2.4500	0.4481
Sample	9	88.05	9.7833	1.1403
Error	9	6.05	.6722	-.1986
Total	19	96.55	-	-

1/ Squared variability divided by degrees of freedom.
2/ $\frac{1}{2}$ log.$_e$ equals $\frac{1}{2}$ log.$_{10}$ times 2.3026, or log.$_{10}$ times 1.1513. These values were calculated by obtaining the products of 2.3026 and one-half the common logarithms of the numbers in column 4. See footnote 2, table 4.

Column 5 of the table shows the one-half natural logarithms of the averages in column 4. For the bias measure of 2.45, containing 1 part of error (0.6722), the $\frac{1}{2}$ $\log._e$ value is 0.4481, and for error alone the $\frac{1}{2}$ $\log._e$ value is -0.1986. The difference between 0.4481 and -0.1986 is 0.6467, which is the z value corresponding thereto, or to bias free from the 1 part of error. To avoid negative logarithms, this difference can be calculated in this instance by moving the decimals in column 4 one place to the right.

In the table showing 5 percent points of the distribution of z, with n_1 equaling 1 and n_2 equaling 9, the z value is 0.8163, indicating that a value of z as great as or greater than 0.8163 would be expected to be obtained by chance alone in not more than 5 percent of the number of cases.[10/] It is apparent, then, that the z value coinciding with the 5 percent point in the distribution of z values is the equivalent of the calculated odds of 19 to 1.

Whenever a derived z value is smaller than that occurring at the 5 percent point, it would seem logical to conclude that the odds

---

10/ Fisher, R. A., Statistical Methods for Research Workers, fourth edition, 1932, table VI, pages 224 and 225. See pages 226 and 227 for 1 percent points of the distribution of z. Readers who wish to estimate the reliability of very small samples may find it convenient to refer to a table prepared by "Student" (Metron V., No. 3, 1925, pp. 105-112.)

The necessity for removing effects of correlation in paired data is emphasized by Dr. W. B. Kemp in an enlightening paper entitled "The Reliability of a Difference Between Two Averages" (Jour. Amer. Soc. of Agronomy, Vol. 16, No. 6, June, 1924, pp. 359-362.) Correlation relative to formulas for error is discussed by "Student" in a paper entitled "On Testing Varieties of Cereals" (Biometrika, Vol. 15, parts 3 and 4, December, 1923, pp. 271-293), and F. D. Rickey, in a paper on "Adjusting Yields to Their Regression on a Moving Average as a Means of Correcting for Soil Heterogeneity" (Jour. Agri. Research, Vol. 27, No. 2, January 12, 1924, pp. 79-90) discusses correlation in paired data.

are less than 19 to 1 that the difference being considered is significant, or less than 19 to 1 against the difference being due to chance alone. Since our calculated z value is only 0.6467, which is appreciably less than the 5 percent value, the magnitude of the difference considered is not regarded as being significant from this standpoint. As will be indicated later, however, considerable importance might be attached to the fact that the differences occur, and the existence of bias, or net difference, might be of special interest in a further study of the possible causal relationship between gin compression and staple-length designation of cotton samples.

Calculating the t value by the formula $t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_d^2}{n}}}$, in which $\bar{x}$ and $\bar{y}$ are the means of the x and y series in table 1, and in which correction is made for degrees of freedom in deriving the standard deviation squared of the differences between paired observations, 1.9094 is obtained, which is less than the value indicating significance in Fisher's table of t. 11/

If the z value occurred far beyond the 5 percent point in the table of z value distributions, the logical conclusion would be that the difference under consideration is markedly significant. According to the distribution of z values, it would be expected that the z values in 95 percent of the cases would be less than 0.8163 when n_1 equals 1 and n_2 equals 9. It is seen, therefore, that the calculated z value of 0.6467 would be expected to occur in this 95 percent group.

11/ Fisher, R. A., Statistical Methods for Research Workers, fourth edition, 1932, table IV.

Another table, which has already been referred to in footnote 10, shows 1 percent points of the distribution of z. With n_1 equaling 1 and n_2 equaling 9, the 1 percent point is found to be 1.1786, indicating that a z value as great as or greater than 1.1786 would be expected to be obtained by chance alone in not more than 1 percent of the cases, and that 99 percent of the z values would be expected to be smaller than the value of 1.1786 shown in the table. If a derived z value corresponds to that shown at the 1 percent point, the odds are 99 to 1 that the difference being considered is significant. If the derived z value is greater than that occurring at the 1 percent point, the odds are more than 99 to 1 that the difference is significant. The lack of significance of such differences does not necessarily mean, of course, that no importance is to be attached to differences in contributions made to total squared variability from the different sources detected.

The use of the tables of z distributions must proceed in this analysis with a complete understanding that n_1 has reference to the larger measure of average squared variability, or variance, and that n_2 refers to the smaller measure. It must be realized also that when the number of observations is small, the difference between calculated measures of average squared variability may be quite large before they become statistically significant, and that as the number of degrees of freedom increases smaller differences may be expected to indicate significance. This applies also, of course, to certain common probability measures.

It will be observed by reference to column 4 of table 2 that the measure for bias is 2.4500; and that for error it is 0.6722. Column 2

shows the corresponding number of degrees of freedom. The measure for "sample" is 9.7833, with 9 degrees of freedom, but the problem of determining the probability of significance of differences has been concerned only with the measures for bias and error. If the total squared variability were separated into only two component parts, we would have 1 degree of freedom for the difference between series and 18 for differences within series.

In column 3 of table 2, as already indicated, 1 part of error, 0.6722, may be understood as being contained in the measure 2.45 in column 3, and 9 parts in the measure 88.05. To obtain percentages representing estimates of the proportionate contributions made from the different detected sources of variability it is first necessary to free these measures from error. The total, 96.55, is then divided into the recalculated parts of variability. Thus we have 96.55 divided into 1.78, [or 2.45 minus (1 times 0.6722)], 82.00, [or 88.05 minus (9 times 0.6722)], and 12.77, [or 6.05 + 9 times 0.6722 plus 1 times 0.6722)], to obtain the proportionate contributions made to total squared variability by bias, sample, and error. The percentages obtained by this procedure are 1.84 for bias, 84.93 for sample, and 13.23 for error. In respect to magnitude of contributions made to total squared variability, error is much more important than bias, indicating that the inconsistent "spread" in the classification of samples represented in table 1 contributed more to total squared variability than did the net difference between the two series of staple-length designations.

It has already been observed that about 95 percent of the z values pertaining to such populations or universes for which n_1 is 1

and n_2 is 9 may be expected to be smaller than the value of 0.8163 shown in the z table referred to. Therefore, the difference between the variances of 2.4500 and 0.6722 designated as bias and error, respectively, in table 2 is not considered significant when interpreted from the table inasmuch as the calculated z value of 0.6467 is less than the z value coinciding with the 5 percent point.12/

12/ The analysis of variance method is admirably adaptable to the study of bias and error in paired and replicate classifications of identical samples by any number of classers, and it has been found useful for this purpose. It is the only method yet suggested that provides for the determination of significance of differences between variances attributable to bias and error, or to any other two sources, and at the same time makes it possible by a modification of the analytical approach to provide a measure of total squared variability and measures of component parts, free from estimated error, that may for some purposes be logically compared with one another. A method suggested by O. T. Weaver provides, however, for the elimination of the effects of modal tendencies of errors of observation (See footnote 5.)

~~at least so far as apparent tendencies are concerned, even though the illustration does represent only a small number of samples.~~

The difference, as already observed, between the means of the x and y series of observations in table 1 is 0.7, representing the extent in sixteenths of an inch to which the y series is greater, on the whole, than the x series. This difference, as indicated, is the same as the average of the algebraic summation of deviations of y observations from the x observations. A principal advantage in separating into its component parts the total summation of squared deviations from the common mean, which has already been illustrated, is that the derived measures readily lend themselves to comparison and interpretation, which may not be possible if the calculated measures are in unlike terms.

The difference of 0.7 between the two means is obviously not directly comparable with the average of squared differences between paired observations, nor is the average difference between the two means comparable with the average of squared deviations from the common mean of the two series. This is attributable not only to the fact that this difference between the means of the two series is in linear terms and cannot be directly compared, therefore, with squared measures, but also, and perhaps chiefly, to the fact that the differences between paired observations contribute the entire "spread" or dispersion, which represents both the bias and the error, the latter having reference to that part of the "spread" or dispersion which is in addition to the bias, or net difference between the series.

In order to show results of the analysis of variance when applied to a larger group of data than that presented in table 1, and in order to further illustrate the practical application of the analytical technique, two additional series containing much greater numbers of observations are presented in table 3. The basic data contained in the first two columns of this table represent the actual classifications of press-box and cut samples of cotton taken from the same bales and classed by the same classers and under as nearly uniform conditions as practicable to maintain at the time.

Table 3. - Distribution of staple-length designations of 16,977 paired press-box and cut samples of cotton

Staple-length designation 1/ (sixteenths of an inch) x 2/	y 3/	x + y	Frequency (n) 4/	xn	yn	(x + y)n
13	13	26	499	6,487	6,487	12,974
13	14	27	742	9,646	10,388	20,034
13	15	28	75	975	1,125	2,100
13	16	29	12	156	192	348
14	13	27	303	4,242	3,939	8,181
14	14	28	5,959	83,426	83,426	166,852
14	15	29	1,744	24,416	26,160	50,576
14	16	30	179	2,506	2,864	5,370
14	17	31	18	252	306	558
14	18	32	2	28	36	64
15	13	28	14	210	182	392
15	14	29	1,177	17,655	16,478	34,133
15	15	30	2,612	39,180	39,180	78,360
15	16	31	636	9,540	10,176	19,716
15	17	32	25	375	425	800
15	18	33	2	30	36	66
16	13	29	4	64	52	116
16	14	30	80	1,280	1,120	2,400
16	15	31	420	6,720	6,300	13,020
16	16	32	1,814	29,024	29,024	58,048
16	17	33	649	10,384	11,033	21,417
16	18	34	11	176	198	374
		Total	16,977	246,772	249,127	495,899
		Mean 5/	-	14.536	14.674	14.605 6/

1/ Selected from unpublished data. Press-box samples are those taken from the gin press-box while the bales were in process of being ginned, whereas cut samples are those taken from the same bales after they had been pressed and tied.

2/ Representing staple-length designations of the press-box samples. (See footnote 1.)

3/ Representing staple-length designations of the cut samples. (See footnote 1.)

4/ The figures in this column represent the number of times each pair of observations occurs. For example, the paired observations of 13 in the x column and 13 in the y column occur 499 times. The total of 1,328

(continued on next page)

4/ (continued) (sum of 499, 742, 75, and 12) designated as 13/16-inch (1st and 4th columns) according to classification of the press-box samples was distributed among 4 staple lengths (2nd and 4th columns) according to classification of the cut samples. Distributions of other lengths are shown by the arrangement of paired classifications.

5/ Calculated for comparative purposes on the lower limits of the staple-length groups. Means based on midpoints may be obtained by adding 0.500.

6/ Representing the common mean of the two series of observations.

1. Correction factor = 495,899 x 14.605 = 7,242,605

2. Total squared variability = 3,599,830 + 3,670,767 - 7,242,605 = 27,992

3. Bias, or x - y squared variability, = $\frac{(246,772)^2}{16,977} + \frac{(249,127)^2}{16,977} -$ 7,242,605 = 175

4. "Sample" variability = $\frac{14,533,631}{2}$ - 7,242,605 = 24,211

5. Error = 27,992 - 24,211 - 175 = 3,606

Having the total squared variability separated into designated component parts, each containing as many parts of average squared variability for error as are indicated by the corresponding degrees of freedom (Column 2, table 4), it is now possible to proceed with the determination of whether or not one estimate of squared variability obtained from n_1 degrees of freedom differs significantly from another estimate of squared variability obtained from n_2 degrees of freedom. This is done, just as in the case of calculations represented in table 2, by deriving the z value equal to half the difference between the natural logarithms of the two measures of average squared variability, or to the difference between the natural logarithms of the corresponding standard deviations (i.e., the difference between the natural logarithms of the square root of the measures of average squared variability). The values in column 4 of table 4 are measures of average squared variability and they may be referred to conveniently as variances. Values in column 5 are the products obtained by multiplying the values in column 4 by 10, which is the equivalent of moving the decimal points one place to the right.

Table 4. - Squared variability contributed from specified sources

1	2	3	4	5	6
Source of squared variability	Degrees of freedom	Squared variability (summation of squares)	Average squared variability 1/	Average squared variability times 10 2/	$\frac{1}{2}$ log.$_e$ 3/
Bias.........	1	175	175.00	1750.00	3.7337
Sample.......	16,976	24,211	1.43	14.30	1.3301
Error........	16,976	3,606	.21	2.10	.3710
Total.....	33,953	27,992	---	---	---

1/ Squared variability divided by degrees of freedom.
2/ Decimals moved one place to the right to avoid negative logarithms in the calculation of values in column 6. Note that in table 2 the negative logarithm of average squared variability for error was used. A paper entitled "Negative Logarithms," and dated May 13, 1932, has been prepared by Norma L. Goudy. This paper is available in mimeographed form.
3/ $\frac{1}{2}$ log.$_e$ equals $\frac{1}{2}$ log.$_{10}$ times 2.3026, or log.$_{10}$ times 1.1513. These values were calculated by obtaining the products of 2.3026 and one-half of the common (five-place) logarithms of the numbers in column 5.

The z value is the difference between any two of the one-half natural logarithms in column 6. Since the problem is concerned first with the difference between bias and error, the desired z value is calculated by subtracting 0.3710 from 3.7337, which leaves 3.3627. In the table showing 5 percent points of the distribution of z, 13/ with n_1 equaling 1 and n_2 equaling infinity, the z value is 0.6729, indicating

13/ Fisher, R. A., Statistical Methods for Research Workers, fourth edition, 1932, table VI, pages 224 and 225. See pages 226 and 227 for 1 percent points of the distribution of z. (Refer to footnotes 8 and 9.)

that a value of z as great as or greater than 0.6729 would be expected to be obtained by chance alone in not more than 5 percent of the number of cases. With n_1 equaling 1 and n_2 equaling 60, the z value occurring at the 5 percent point is 0.6933. In the table showing 1 percent points of the distribution of z, with n_1 equaling 1 and n_2 equaling infinity, the z value is 0.9462. It is obvious, therefore, that the z value of 3.3627 indicates the difference between the two measures for bias and error in table 4 to be very highly significant.

In order to obtain percentages representing proportionate contributions made to total squared variability from the different detected sources, it is first necessary to free the designated component parts of squared variability from error. The separation of error from the designated parts of squared variability other than error is accomplished by obtaining the products of the average squared variability measure for error and the degrees of freedom, and then making the proper subtractions.

Numerical measures of average squared variability are expressed in only two decimals in column 4 of table 4. These measures have been recalculated and carried to four decimals. They are presented in column 4 of the following table, the first three columns of which were adapted from table 4. Following the table are calculations indicating the procedure by which error is separated from the measures in column 3.

Table 5. - Sources and measures of squared variability

1	2	3	4
Source of squared variability 1/	Degrees of freedom 1/	Squared variability 1/ (summation of squares)	Average squared variability
Bias.............	1	175	175.0000
Sample...........	16,976	24,211	1.4262
Error............	16,976	3,606	.2124
Total.........	33,953	27,992	-

1/ Adapted from table 4.

Bias, free from error, = 175 minus (1 times 0.2124) = 174.7876

Sample variability, free from error, = 24,211 minus (16,976 times 0.2124) = 20,605.2976

Error = 3606 plus (16,976 times 0.2124) plus (1 times 0.2124) = 7,211.9148

Total = 27,992.0000

The proportionate contributions made to the total from each detected source are calculated by dividing 27,992 into each of the derived measures following the table. By this procedure there is obtained 0.63 percent for bias, 73.61 percent for sample, and 25.76 percent for error. When the variability contributed from the different sources is evaluated in this way, the bias is relatively small in magnitude.